Forecasting With Stable Seasonal Pattern Models With an Application to Hawaiian Tourism Data

Rong CHEN
Department of Statistics, Texas A&M University, College Station, TX 77843-3143 (*chen@stat.tamu.edu*)

Thomas B. FOMBY
Department of Economics, Southern Methodist University, Dallas, TX 75275-0491 (*tfomby@mail.smu.edu*)

We propose several variations of the stable seasonal pattern (SSP) model first introduced by Marshall and Oliver and study their prediction procedures. Depending on the type of data (count data or continuous variable), we propose different treatments. Previously SSP models have been applied to trendless data or adapted to trending data in ad hoc ways. In the models considered here, conditional independence allows the seasonal pattern and trend to be modeled separately, whereas prediction uses both efficiently. In an out-of-sample forecasting experiment conducted on Hawaiian tourism data, one of the proposed variations demonstrates its long-term forecasting potential relative to seasonal Box–Jenkins autoregressive integrated moving average and transfer function models.

KEY WORDS: Conditional Poisson model; Forecasting; Seasonality.

The analysis of seasonal time series has a long and storied history beginning with Gilbart (1854, 1856, 1865). See Hylleberg (1992) for an extensive historical review. Of course, there exist many characterizations of seasonality in time series data. They range from stochastic seasonality as discussed by Box and Jenkins (1976) to the unobserved-components seasonality of Harvey (1989) to deterministic seasonality vis-à-vis dummy variables in multiple regression as given by Johnston (1984) to stable seasonal pattern models as given by Oliver (1987).

This article is concerned with the extension of the latter literature. The stable seasonal pattern (SSP) model was first developed in longitudinal data analysis of student enrollment and attendance at a large university (Marshall and Oliver 1970, 1979). Moreover, an early Bayesian approach was used by Chang and Fyffe (1971) to forecast the sales of seasonal style goods. A more extensive Bayesian approach was introduced by Oliver (1987). In many applications the fraction of the number of events that occur within a season is stable over time, even though the end-of-season totals and the cumulative count of events during a season may be unknown (e.g., Marshall and Oliver 1970; Chang and Fyffe 1971; Green and Harrison 1973; Grinold and Marshall 1977). This problem is of interest in many applications, particularly to the tourism industry, in which seasonal patterns almost always exist and prediction of the end-of-season total is important to inventory management, marketing strategy, and many other decision-making processes involving local governments and businesses affected both directly and indirectly by the tourism industry. Other potential applications of the proposed SSP method include financial budgeting, profit forecasting, and water resources and energy consumption management, in which the prediction of the end-of-year total is important.

In this article we propose several variations of the SSP model and study their prediction procedures. Depending on the type of data observed (count data or continuous variable), we propose different treatments. Note that, when the value range of the count data is large, most of the time it should be treated as a continuous variable.

Our proposed methods are applied to Hawaiian tourism data obtained from the Hawaii Visitors Bureau. Hawaii is a major worldwide tourism site. Approximately 40% of the state's gross domestic product is either directly or indirectly dependent on tourism. Hence, for the businesses and local governments of Hawaii the prediction of end-of-season totals and cumulative within-season amounts of tourism is very important. There are two major sources of tourists visiting Hawaii each year, the westbound tourists originating from the United States, Canada, and Europe and the eastbound tourists originating from Japan, Australia, and other Pacific countries. Figure 1(a) shows the monthly observations from January 1970 through December 1995. Figure 1(b) shows the aggregated yearly totals for the same period. Following Berman and Shumway (1998), in Figure 2(a) we show the perspective plot of the ratios of the monthly observation to its respective yearly total. Clearly we see a relatively stable seasonal pattern, with March, December, and the summer being the heaviest traveled months. This stable pattern is the basis for our proposed SSP modeling approaches and prediction procedures. Further support of the SSP model will be given in later sections. The downturn around 1991–1992 is partially due to the recession in the United States around that time (particularly the severe recession in California).

The rest of the article is organized as follows. Section 1 introduces the SSP model in a general setting. In Sections 2 and 3 we examine the SSP model and its prediction

Journal of Business & Economic Statistics
October 1999, Vol. 17, No. 4

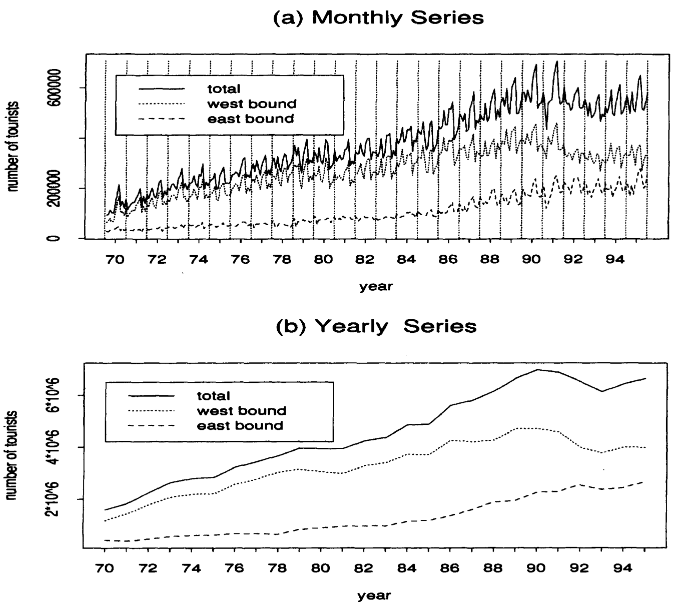

Figure 1. Monthly (a) and Yearly (b) Hawaiian Tourist Series.

procedures under two situations, discrete observations and continuous observations. These procedures are then applied to the Hawaiian tourism data and the results are presented within each section. Because Hawaiian data is in the range of 10^5, treating it as count data may not be appropriate. Here we use it to demonstrate our procedure. A comparison of the procedures with the more conventional Box–Jenkins method of forecasting seasonal data is presented in Section 5 along with a summary of the results of the article.

1. SEASONAL TIME SERIES WITH SSP

Consider a seasonal time series $\ldots, x_{t1}, x_{t2}, \ldots, x_{td}, \ldots,$ where $t = 1, 2, \ldots, n$ are the season indices and d is the length of the season. The season total is then given by $y_t = x_{t1} + \cdots + x_{td}$.

An SSP model assumes that $(x_{t1}, \ldots, x_{td})$ are conditionally independent for $t = 1, \ldots, n$, given the season total y_t. That is,

$$\pi[x_{t_1 1}, \ldots, x_{t_1 d}, \ldots, x_{t_k 1}, \ldots, x_{t_k d} | y_{t_1}, \ldots, y_{t_k}] = \prod_{i=1}^{k} \pi[x_{t_i 1}, \ldots, x_{t_i d} | y_{t_i}],$$

for any set $(t_1, \ldots, t_k)$ where $\pi[\cdot|\cdot]$ denotes the conditional density. This structure allows us to model the series with a two-component model, $\pi[x_{t1}, \ldots, x_{td} | y_t] \sim M_1(y_t, \boldsymbol{\theta})$, and $\{y_t\} \sim M_2(\boldsymbol{\beta})$, where M_1 is a $(d-1)$-variate distribution and M_2 is a time series model for the season total series $\{y_t\}$. Those two models are to be determined or selected for specific applications, and $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are parameters to be estimated. This model structure is quite similar to the state-space models in which the season totals can be viewed as the states and the x's as the observations, as shown in Figure 3. Such a two-component model was used by Burman and Shumway (1998) in modeling seasonal time series in which the yearly trend and the seasonal pattern were modeled separately, using semiparametric methods.

There are several important features of this model. First, it is assumed that the correlation between seasons is only through the season total y_t, not individual observations within the seasons. Second, except for the effect of the season total, different seasons follow the same model, with the same parameters. For example, if the model M_1 is a multinomial distribution, then the proportion $\boldsymbol{\theta}$ is the same for all seasons. Third, in the Hawaiian tourism data, the aggregated season-total series y_t can (and should) be modeled separately, independently of the original series.

Chang and Fyffe (1971) used the SSP model to study the sale of seasonal goods. They assumed $y_t = a + bw_t + \varepsilon_t$ for the season total model M_2, where w_t is some observable exogenous variable. For M_1, they assumed that $x_{ti} = \theta_i y_t + n_{ti}$, where $\sum_{i=1}^{d} \theta_i = 1, 0 < \theta_i < 1$, and $n_{ti} \sim \mathrm{N}(0, \sigma_i^2)$ with $E(n_{ti} n_{tj}) = 0$ for $i \neq j$. Although under this model we have $E(\sum_{i=1}^{d} x_{ti}) = y_t$, we do not have $\sum_{i=1}^{d} x_{ti} = y_t$, which is troublesome. In addition, the assumption that n_{ti} and n_{tj} are uncorrelated implies that x_{ti} and x_{tj} are conditionally uncorrelated, given y_t. This assumption is also questionable. In Section 3 we will introduce a similar model that does not have these problems.

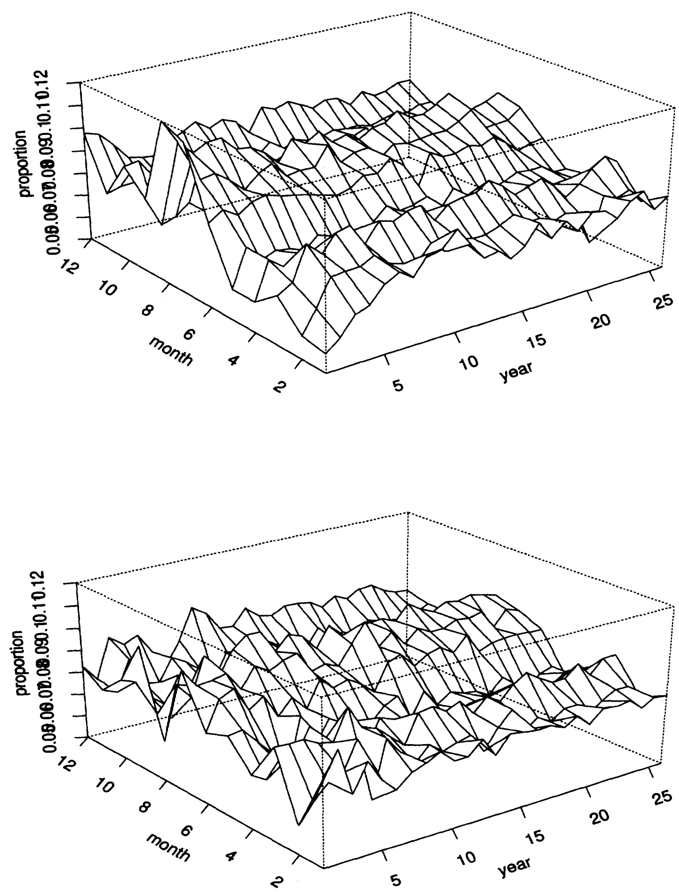

Figure 2. Perspective Plot of Monthly Proportion to Yearly Total: (a) Hawaii Series (top figure); (b) a Simulated Series From the Gaussian-Multinomial Model (bottom figure).

$$
\begin{array}{ccccccc}
 & (x_{t1},\ldots,x_{td}) & & (x_{t+1,1},\ldots,x_{t+1,d}) & & \\
 & \Uparrow (M_1) & & \Uparrow (M_1) & & \\
\cdots \Longrightarrow & y_t & \overset{(M_2)}{\Longrightarrow} & y_{t+1} & \Longrightarrow & \cdots
\end{array}
$$

Figure 3. Illustration of the General SSP Model Structure.

2. FORECASTING WITH THE SSP MODEL: DISCRETE CASE

2.1 General Model Setting and Prediction

When x_{tk} are integer-valued seasonal time series and if we assume the seasonal pattern is stable over seasons, then one obvious choice for model M_1 is $\pi[x_{t1},\ldots,x_{td}|y_t] \sim \text{multinomial}(y_t,\boldsymbol{\theta})$ with $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_d), 0 \le \theta_i \le 1$, and $\sum_{i=1}^d \theta_i = 1$. The next step is to choose a model for the season-total time series. If we can assume that the interarrival times between tourists are independent and exponentially distributed, then we can assume a nonhomogeneous Poisson model for the season total, in which the arrival rate is a function of the lag values of season total and some exogenous variables. Let $\mathbf{y}_t = (y_t, y_{t-1},\ldots,y_1)$ be the lag variables and $\mathbf{z}_t$ be a set of exogenous variables observable at time t. Define a conditional Poisson model as

$$\pi[y_t|\mathbf{y}_{t-1},\mathbf{z}_t,\boldsymbol{\beta}] \sim \text{Poisson}(\lambda_t), \tag{1}$$

where $\lambda_t = f(\mathbf{y}_{t-1},\mathbf{z}_t,\boldsymbol{\beta})$ is a positive function of the past observations or some observable exogenous variables and $\boldsymbol{\beta}$ is the associated parameter vector.

At the kth observation of the tth season, we wish to predict the season total y_t, given $\mathbf{y}_{t-1}$ and $x_{t1},\ldots,x_{tk}, 0 \le k < d$. Note that, under the conditional independence assumption,

$$
\begin{aligned}
&\pi[y_t|x_{11},\ldots,x_{t-1,d},x_{t1},\ldots,x_{tk},\mathbf{y}_{t-1},\mathbf{z}_t,\boldsymbol{\beta},\boldsymbol{\theta}]\\
&= \pi[y_t|x_{t1},\ldots,x_{tk},\mathbf{y}_{t-1},\mathbf{z}_t,\boldsymbol{\beta},\boldsymbol{\theta}]\\
&\propto \pi[x_{t1},\ldots,x_{tk}|y_t,\boldsymbol{\theta}]\pi[y_t|\mathbf{y}_{t-1},\mathbf{z}_t,\boldsymbol{\beta}]\\
&\propto \frac{(1-\sum_{i=1}^k \theta_i)^{y_t}}{(y_t - \sum_{i=1}^k x_{ti})!}\lambda_t^{y_t}\\
&\sim \sum_{i=1}^k x_{ti} + \text{Poisson}\left[\left(1-\sum_{i=1}^k \theta_i\right)\lambda_t\right].
\end{aligned}
$$

Hence, the least squares prediction (the conditional mean) is

$$\hat{y}_t(k) = \sum_{i=1}^k x_{ti} + \sum_{i=k+1}^d \theta_i\lambda_t. \tag{2}$$

Again, $\lambda_t = f(\mathbf{y}_{t-1},\mathbf{z}_t,\boldsymbol{\beta})$ is the prediction of y_t using only the season total series model M_2. The mean squared prediction error is

$$E[(y_t - \hat{y}_t(k))^2|\mathbf{y}_{t-1},\mathbf{z}_t,\boldsymbol{\beta},\boldsymbol{\theta}] = \sum_{i=k+1}^d \theta_i\lambda_t.$$

A closer examination of Equation (2) reveals the effect of the memoryless property of the Poisson distribution. For $k = 0$, where there is no observation in season t, the prediction of y_t is solely from the model of season-total series, which can be partitioned as

$$\hat{y}_t = \lambda_t = \sum_{i=1}^d \theta_i\lambda_t,$$

where each term in the last summation is the predicted contribution to the season total from each observation within the season because $\pi(x_{t1},\ldots,x_{td}|\mathbf{y}_{t-1},\boldsymbol{\theta},\boldsymbol{\beta}) \sim \text{multinomial}(\lambda_t,\theta_1,\ldots,\theta_d)$. Equation (2) shows that, when we start to observe observations within the season, the corresponding estimated contribution (at the beginning of the season) is replaced by the observed value but no adjustment is made to the rest of the predicted contributions. Note that this is only true with the Poisson distribution. Other distributional assumptions and models may not possess such a memoryless property, as we will show in later sections.

Remark The conditional Poisson model in (1) is somewhat restrictive, in a way that the conditional variance of y_t is also restricted to be λ_t. There are many other more flexible choices. For example, it is possible to use the conditional negative binomial model for more dispensed variance structure. Specifically, let

$$
\begin{aligned}
\pi[\mathbf{y}_t|\mathbf{y}_{t-1},\mathbf{z}_t,\beta] &= \text{NegBinomial}(r_t,p_t)\\
&= \binom{y_t+r_t}{r_t} p_t^{r_t}(1-p_t)^{y_t-r_t},
\end{aligned}
$$

where r_t, p_t are functions of $(\mathbf{y}_{t-1},\mathbf{z}_t)$. In this case, the variance of y_t is $\mu_t(1+\delta\mu_t)$, where $\mu_t = r_t(1-p_t)/p_t$ and $\delta = 1/r_t$.

In this case, it is easily shown that

$$
\begin{aligned}
&\pi[y_t|x_{t1},\ldots,x_{tk},\mathbf{y}_{t-1},\mathbf{z}_t,\beta] + r_t\\
&\sim \text{NegBinomial}\left(\sum_{i=1}^k x_{ti} + r_t; 1 - \left(1-\sum_{i=1}^k \theta_i\right)(1-p_t)\right).
\end{aligned}
$$

Hence

$$\hat{y}_t(k) = \left(\sum_{i=1}^k x_{ti} + r_t\right)\frac{(1-\sum_{i=1}^k \theta_i)(1-p_t)}{1-(1-\sum_{i=1}^k \theta_i)(1-p_t)} - r_t.$$

We will use the conditional Poisson model in this article.

2.2 Conditional Poisson Time Series Models for Season Total y_t

To use the model and the prediction procedure in Section 2.1, we need to specify the conditional Poisson model for the season total. Here we introduce some possible models.

2.2.1 Conditional Poisson Autoregressive Models. Autoregressive (AR) models have been used extensively for real-valued time series (Box and Jenkins 1976). Integer-valued time series models were studied by McKenzie (1985, 1988), who proposed the Poisson AR model in which the marginal distribution of the process follows a Poisson distribution and the model can be written in an autoregressive-like form; that is, y_t depends on its lagged variables. The conditional distribution of the Poisson AR model is quite complex, however, and is not suitable for our analysis. Harvey and Fernandes (1989) also proposed several useful time series models for count or qualitative observations, with a two-layer model structure.

Here we introduce a conditional Poisson AR (CPAR) model for the season-total series. Again, let $\mathbf{y}_t = (y_t, y_{t-1}, \ldots, y_1)$. A CPAR($p$) model assumes the form of

$$\pi[y_t|\mathbf{y}_{t-1}, \boldsymbol{\beta}] \sim \text{Poisson}\left(\beta_0 + \sum_{i=1}^{p} \beta_i y_{t-i}\right),$$

where $\beta_0 > 0$ and $\beta_i \geq 0, i = 1, \ldots, p$.

Some probabilistic properties of the model can be easily found. For example, by applying Tweedie's criterion (Tweedie 1975), we obtain that a sufficient condition for the process to be stationary and geometrically ergodic is $\sum_{i=1}^{p} \beta_i < 1$. In particular, for $p = 1$, the mean and variance of the stationary distribution are $\beta_0/(1 - \beta_1)$ and $\beta_0/(1-\beta_1)(1-\beta_1^2)$. Obviously, the marginal stationary distribution is no longer a Poisson distribution. Another reason that the marginal distribution is of less concern is that the season-total series may not be stationary, as we see in the Hawaiian tourist series. Because our prediction procedure requires only the conditional distribution, stationarity (and marginal distribution) is not our main concern and is not important for prediction purposes.

Note that, given $y_1, \ldots, y_n$, the conditional log-likelihood function is

$$l(\boldsymbol{\beta}) = \sum_{t=p+1}^{n} y_t \log\left(\beta_0 + \sum_{i=1}^{p} \beta_i y_{t-i}\right) + \sum_{t=p+1}^{n} \left(\beta_0 + \sum_{i=1}^{p} \beta_i y_{t-i}\right),$$

which can be optimized easily under the restriction that $\beta_0 > 0$ and $\beta_i \geq 0$ for $i = 1, \ldots, p$.

2.2.2 Conditional Poisson AR Models With Exogenous Variables. In many applications, exogenous variables are available for improving model fitting and prediction. Let $\mathbf{z}_t = (z_t, z_{t-1}, \ldots)$ be an exogenous variable series. A CPAR model with exogenous variables [CPAR(p)x] can be defined as $\pi[y_t|\mathbf{y}_{t-1}, \mathbf{z}_t] = \text{Poisson}(\lambda_t)$, where

$$\lambda_t = a_0 + \sum_{i=1}^{p} a_i y_{t-i} + \sum_{i=0}^{q} b_i (z_{t-i-l} + C)$$

for $a_0 > 0, a_i \geq 0$, and $b_i \geq 0$. The constant C is used so that $z_t + C \geq 0$ for all t and can be viewed as the intercept term in the generalized linear model. Note that the preceding setting is proper only when the y series and the z series are positively correlated. To incorporate a negatively correlated exogenous variable, one can impose restrictions that $b_i \leq 0$ and $z_t + C \leq 0$.

The parameter l is the lag time for the changes in the exogenous variable to have an effect on y_t. For our prediction purpose, we need $l > 0$ to avoid predicting the value of the exogenous variable for the current season. Again, the conditional log-likelihood function of the model can be easily obtained and optimized.

2.2.3 Remark. If, instead of specifying a parametric model M_2 for the season total series $\{y_t\}$, we simply assume that the y_t are independent, with an improper uniform distribution, then we can obtain the following naive predictors:

$$\pi[y_t|x_{t1}, \ldots, x_{tk}, \boldsymbol{\theta}] \sim \sum_{i=1}^{k} x_{ti} + \text{NegBinomial}\left(\sum_{i=1}^{k} x_{ti}; \sum_{i=1}^{k} \theta_i\right).$$

Hence, the least squares predictor (the conditional mean) is

$$\hat{y}_t(k) = \frac{\sum_{i=1}^{k} x_{ti}}{\sum_{i=1}^{k} \theta_i}, \tag{3}$$

with prediction variance

$$\text{var}[y_t|x_{t1}, \ldots, x_{tk}] = \frac{(\sum_{i=1}^{k} x_{ti})(1 - \sum_{i=1}^{k} \theta_i)}{(\sum_{i=1}^{k} \theta_i)^2}.$$

When the size of the seasonal-total series is small and there is no clear model structure for M_2, the independence assumption may be useful, as shown by Marshall and Oliver (1970) for student-enrollment data.

2.3 Hawaiian Tourism Series as Count Data

In this section we apply the discrete SSP model to the Hawaiian tourism series. As mentioned before, count numbers in the range of 10^6 might be better treated as continuous observations. But here we use it as an example of the discrete SSP model.

In this section and all the sections concerning the analysis results of the Hawaiian tourist series, we will focus on out-of-sample prediction performance. For this purpose, we use the data from 1970 to 1984 for model selection. Starting in January 1985, when each monthly observation is added to the series, we reestimate the model and obtain the prediction of the corresponding yearly total, resulting in 12 predictions for each year. Those predictions are then compared with the true annual data.

We have entertained several models for the yearly total series. Due to the limited sample size of the annual series, only CPAR models of order 1 and 2 are considered. It appears that the CPAR(2) model is reasonable, with lower Bayesian information criterion (BIC). In general, model-selection procedures such as the information criteria can be used. Actually, because our main interest is in the prediction of the yearly total, the rolling out-of-sample prediction performance we present here is a reasonable and useful criterion itself. In Figure 4, the x's show the "rolling" prediction result for years 1985 to 1995 using the CPAR(2) model for M_2 and multinomial distribution for M_1. The straight lines are the true yearly total.

Because Hawaii's tourist industry is closely related to the visiting nations' economic conditions, we use some economic indexes to improve our prediction performance. As mentioned earlier, tourists visiting Hawaii consist of two major groups, the westbound tourists, mainly from

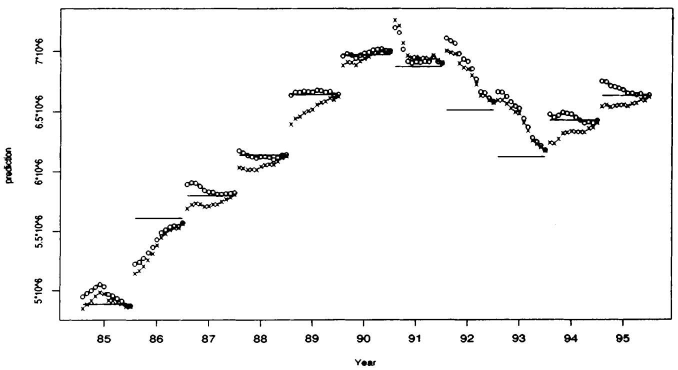

Figure 4. Prediction Using the CPAR Model. The straight lines are the true yearly totals. The x's are predictions using a CPAR(2), model, and the o's are predictions using a CPAR(1)x model.

the United States, and the eastbound tourists, mainly from Japan. Because the two regions can have quite different economic statuses each year, we model the annual series of the two groups separately, using their corresponding economic indexes. Let y_t^w be the yearly total of westbound tourists and y_t^e be that of eastbound tourists. We use P_t^w, the real personal disposable income (PDI) of the United States, and P_t^e, the real PDI of Japan, as our exogenous variables.

Specifically, we use the following CPAR(1)x model. Assume that $y_t^w \sim \text{Poisson}(\lambda_t^w)$ and $y_t^e \sim \text{Poisson}(\lambda_t^e)$, where $\lambda_t^w = a_1 + a_2 y_{t-1}^w + a_3 P_{t-1}^w$ and $\lambda_t^e = b_1 + b_2 y_{t-1}^e + b_3 P_{t-1}^e$. In addition, we also assume that y_t^w and y_t^e are conditionally independent, given y_{t-1}^w and y_{t-1}^e. Then it is easy to show that $\pi[y_t | \mathbf{y}_{t-1}^w, \mathbf{y}_{t-1}^e] = \pi[y_t^w + y_t^e | \mathbf{y}_{t-1}^w, \mathbf{y}_{t-1}^e] \sim \text{Poisson}(\lambda_t^w + \lambda_t^e)$. Table 1 shows the maximum likelihood estimates of the parameters in the multinomial distribution and the preceding CPAR(1)x model using all of the data.

In Figure 4, the o's are the predictions using the CPAR(1)x model. We can see that the prediction at the beginning of the year using the model with exogenous variables is closer to the true value, except at the turning point. Because our yearly-total series is relatively short, we did not attempt to include more lags. Other exogenous variables, such as California nonagricultural employment, U.S. nonagricultural employment, the dollar/yen exchange rate, and Japanese industrial employment, were also investigated but are not reported here.

In the preceding analysis, we assume that the seasonal pattern of the two groups is the same. Of course, we can model the seasonal pattern separately, though our own analysis has shown that the difference is small.

As a goodness-of-prediction test, we also obtained the prediction of a simple model that assumes that all θ_i's are equal to $\frac{1}{12}$. The performance of our model is significantly better.

3. FORECASTING WITH THE SSP MODEL: CONTINUOUS CASE

3.1 General Model Setting and Prediction

To begin this section, we first define the *Gaussian-multinomial* (G-MN) *distribution*. A d-dimensional random variable $(x_1, \ldots, x_d)$ is said to follow G-MN $(y, \theta_1, \ldots, \theta_d, \sigma^2)$ if $\pi[x_1, \ldots, x_{d-1}] \sim \text{N}(\mu, \sigma^2 \Sigma)$ and

$$x_d = y - \sum_{i=1}^{d-1} x_i,$$

where

$$\mu = (\theta_1 y, \ldots, \theta_{d-1} y)'$$

$$\Sigma = \begin{pmatrix} \theta_1(1-\theta_1) & -\theta_1\theta_2 & \cdots & -\theta_1\theta_{d-1} \\ -\theta_1\theta_2 & \theta_2(1-\theta_2) & \cdots & -\theta_2\theta_{d-1} \\ \cdots & \cdots & \cdots & \cdots \\ -\theta_1\theta_{d-1} & -\theta_2\theta_{d-1} & \cdots & \theta_{d-1}(1-\theta_{d-1}) \end{pmatrix}$$

for $0 < \theta_i < 1$ and $\sum_{i=1}^d \theta_i = 1$.

This distribution can be viewed as a continuous version of the multinomial distribution. The mean and correlation coefficient matrix are the same as the multinomial distribution, except for the extra variance parameter σ^2. In the multinomial distribution, the variance varies with the seasonal total, though in a very restricted way. In the G-MN distribution, the variance is constant. It is possible to allow for varying variance (depending on observable variables) similar to the weighted regression setting. Here we choose to assume constant variance.

It can be shown that the G-MN distribution also processes the combination property of the multinomial distribution. For example, if $(x_1, \ldots, x_d) \sim \text{G}-\text{MN}(y, \theta_1, \ldots, \theta_d, \sigma^2)$, then $((x_1 + x_2), x_3, \ldots, x_d) \sim \text{G}-\text{MN}(y, (\theta_1 + \theta_2), \theta_3, \ldots, \theta_d, \sigma^2)$. More importantly, this property implies that, for $1 \le k < d$,

$$\sum_{i=1}^k x_i \sim N(\gamma_k y, \sigma^2 \gamma_k (1 - \gamma_k)),$$

where $\gamma_k = \sum_{i=1}^k \theta_i$. This feature allows us to estimate γ_k much more easily than estimating $\theta_1, \ldots, \theta_d$ in high-dimensional space. Our prediction procedure only requires the estimate of γ_k. In fact, it can be easily shown that the maximum likelihood estimator of $\gamma_k = \sum_{i=1}^k \theta_i$ is

$$\hat{\gamma}_k = \frac{\sum_{t=1}^n (\sum_{i=1}^k x_{ti}) y_t}{\sum_{t=1}^n y_t^2} \quad (4)$$

with standard error $\hat{\sigma}(\hat{\gamma}_k(1 - \hat{\gamma}_k)/\sum y_t^2)^{1/2}$. The maximum likelihood estimator for σ^2 is $\hat{\sigma}^2 = \sum_{t=1}^n (\mathbf{x}_t -$

Table 1. Maximum Likelihood Estimates (standard errors) of the Parameters in the Discrete SSP Model for the Hawaii Series Using All of the Data

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}	θ_{11}	θ_{12}	a_1	a_2	a_3	b_1	b_2	b_3
Estimate	.0810	.0784	.0875	.0777	.0763	.0879	.0938	.0981	.0750	.0790	.0782	.0872	12,764	.782	27,490	399,865	.889	24.08
Std. err.	.000025	.000025	.000026	.000025	.000024	.000026	.000027	.000027	.000024	.000025	.000025	.000025	292.7	.00015	14.4	380.3	.000024	.046

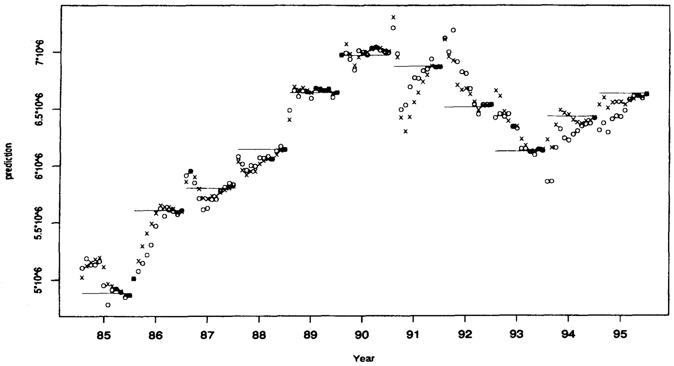

Figure 5. Prediction Using Gaussian AR Model and ARIMA Model. The straight lines are the true yearly totals. The x's are predictions using an SSP model with AR(1) for the annual series, and the o's are predictions using an ARIMA model.

$\hat{\boldsymbol{\theta}} y_t)'\hat{\Sigma}^{-1}(\mathbf{x}_t - \hat{\boldsymbol{\theta}} y_t)/n$.

If, in addition, we assume that the season total is also a Gaussian series, $\pi[y_t|\mathbf{y}_{t-1}] \sim \mathrm{N}(\mu_t, \sigma_y^2)$, where $\mu_t = f(\mathbf{y}_{t-1}, \boldsymbol{\beta})$, then forecasting of y_t, given $y_{t-1}, \ldots, y_1$ and $x_{t1}, \ldots, x_{tk}$ for $0 \le k < d$, can be easily done as follows.

Let $\mathbf{x}_t^{(k)} = (x_{t1}, \ldots, x_{tk})', \boldsymbol{\theta}_k = (\theta_1, \ldots, \theta_k)'$, and $\Sigma_k = (v_{ij})_{k\times k}$, where $v_{ij} = -\theta_i\theta_j$ for $i \ne j$ and $v_{ii} = \theta_i(1-\theta_i)$ and $i, j \le k$. Then

$$\begin{aligned}\pi[y_t|x_{t1}, \ldots, x_{tk}, \mathbf{y}_{t-1}] &\propto \pi[x_{t1}, \ldots, x_{tk}|y_t]\pi[y_t|\mathbf{y}_{t-1}] \\ &\sim N(\mu_{t,k}, \sigma_k^2),\end{aligned}$$

where

$$\sigma_k^2 = \left(\frac{\boldsymbol{\theta}_k'\Sigma_k^{-1}\boldsymbol{\theta}_k}{\sigma^2} + \frac{1}{\sigma_y^2}\right)^{-1} = \left(\frac{\gamma_k}{(1-\gamma_k)\sigma^2} + \frac{1}{\sigma_y^2}\right)^{-1}$$

and

$$\begin{aligned}\mu_{t,k} &= \left(\frac{\mathbf{x}_t^{(k)'}\Sigma_k^{-1}\boldsymbol{\theta}_k}{\sigma^2} + \frac{\mu_t}{\sigma_y^2}\right)\sigma_k^2 \\ &= \left(\frac{\sum_{i=1}^k x_{ti}}{(1-\gamma_k)\sigma^2} + \frac{\mu_t}{\sigma_y^2}\right)\sigma_k^2. \qquad (5)\end{aligned}$$

Hence, the least squares prediction is $\mu_{t,k}$ in (5). Unlike the Poisson case in which observations within the prediction season are only used to correct the prediction of that observation, here the new observation has an effect on the entire prediction. In fact, (5) can be rewritten as

$$\begin{aligned}\mu_{t,k} &= \frac{1}{\gamma_k + (1-\gamma_k)c}\sum_{i=1}^k x_{ti} + \frac{c}{\gamma_k + (1-\gamma_k)c}\sum_{i=k+1}^d \theta_i\mu_t \\ &= w_{1k}\frac{\sum_{i=1}^k x_{ti}}{\gamma_k} + w_{2k}\frac{\sum_{i=k+1}^d \theta_i\mu_t}{1-\gamma_k}, \qquad (6)\end{aligned}$$

where $c = \sigma^2/\sigma_y^2$. Here, $\sum_{i=1}^k x_{ti}/\gamma_k$ and $\sum_{i=k+1}^d \theta_i\mu_t/(1-\gamma_k)$ are the prediction of the yearly total based on only the monthly G-MN model and only the annual Gaussian model, respectively. The weights $w_{1k} = \gamma_k/(\gamma_k + (1-\gamma_k)c)$ and $w_{2k} = (1-\gamma_k)c/(\gamma_k + (1-\gamma_k)c)$ dictate the contribution of the monthly G-MN model and the annual Gaussian model to the prediction of the yearly total.

Note that, when the model for the season-total series and the model for observations within the season have equal precision—that is, $\sigma^2 = \sigma_y^2$—then $\mu_{t,1} = \sum_{i=1}^k x_{ti} + \sum_{i=k+1}^d \theta_i\mu_t$. Hence, the adjustment procedure is simply to replace the predicted contribution of the individual observations by their observed values, which is what we have seen in the Poisson model. When $c < 1$, however, then $\sum_{i=1}^k x_{ti}$ bears more weight in the adjustment procedure, because the individual observation has less variance (more information) than the season-total model. On the other hand, if the individual observation is more volatile and the season total is more stable, then the adjustment procedure puts less weight on the individual observation within the season.

Many possible Gaussian models can be used for the season-total series. In fact, if we assume linearity, all the tools of Box and Jenkins (1976) can be used to model the series. For example, the Gaussian random walk, $\mu_t = y_{t-1}$, the Gaussian AR(p) model, $\mu_t = \phi_0 + \sum_{i=1}^p \phi_i y_{t-i}$, the Gaussian AR($p$) model with time trend, $\mu_t = a + bt + \sum_{i=1}^p \phi_i y_{t-i}$, and the AR(1) model with exogenous variable, $\mu_t = \phi_0 + \phi_1 y_{t-1} + \theta z_{t-1}$, are all possible models for the season-total series. This is the advantage of this continuous setting because Gaussian time series analysis has a more mature literature and there are many well-tested techniques that can be used.

Remark Again, if we simply assume that the y_t are independent, with an improper uniform distribution, then the naive predictor in the continuous case becomes

$$\pi[y_t|x_{t1}, \ldots, x_{tk}] \sim \mathrm{N}\left(\frac{\sum_{i=1}^k x_{ti}}{\gamma_k}, \frac{\sigma^2(1-\gamma_k)}{\gamma_k}\right).$$

Hence, the naive predictor is the same as that in the count-data case (3), but the prediction error is different.

3.2 Hawaiian Tourist Series as Continuous Variables

Figure 5 presents the prediction results for a Gaussian model of the Hawaiian tourist series. The figure reads in the same way as Figure 4. The x's are the predictions of using the prediction procedure described in the preceding section, with Gaussian AR(1) for the annual series. This model was chosen with both Akaike's information criterion (AIC) and BIC from a large set of possible models for the annual series, with or without the exogenous variables mentioned in Section 2.3, with both aggregated or disaggregated (westbound and eastbound) series. From the figure, we can see that, at the beginning of the year, the Gaussian AR model misses the downturn around 1990 badly, compared to that of the Poisson AR model. When we start to observe partial monthly observations and make corrections of our predicted total based on this new information, however, the Gaussian model seems to work better overall. Table 2 shows the maximum likelihood estimates of the parameters in the G-MN distribution and the Gaussian AR(1) model using all of the data. Note that the annual model has larger variance σ(AR)

Table 2. Maximum Likelihood Estimates (standard errors) of θ in the Gaussian-Multinomial Model Using All of the Data

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}	θ_{11}	θ_{12}	σ(G-MN)	ϕ_0	ϕ_1	σ(AR)
Estimate	.0815	.0782	.0875	.0778	.0768	.0878	.0935	.0974	.0754	.0785	.0785	.0870	64,730	402,612	.9549	247,115
Std. err.	.00072	.00070	.00074	.00070	.00070	.00074	.00076	.00078	.00069	.00071	.00071	.00071		1.43E10	.03	

than the G-MN model σ(G-MN). Hence, the correction procedure puts more weight on the monthly observations.

We have performed a parametric bootstrap procedure (Tsay 1992) to check if the G-MN distribution assumption (i.e., the SSP assumption) holds. Using the estimated parameters and true annual totals, we simulated 1,000 series of the same size as our Hawaiian series, using the G-MN model. Figure 2(b) is the perspective plot of monthly proportions of one randomly selected bootstrap series. We see that it is very similar to Figure 2(a). With the bootstrap series, we obtained the bootstrap distribution of the mean, variance, skewness, kurtosis, median, and interquartile range of the monthly proportions for each month (i.e., for each series we have 72 statistics). Among the 72 true values of these statistics from the original Hawaiian data, 8 were out of the 95% bootstrap confidence interval. The month of August contributed 4 of those "unusual" cases (mean, variance, skewness, and kurtosis). The median and interquartile range of August are comfortably inside the bootstrap distribution range. This procedure partially reinforces the G-MN model and the SSP model.

4. COMPARISON AND SUMMARY

In this section we present a comparison of the prediction performances of the different procedures we have proposed. We also compare our models to a standard linear seasonal autoregressive integrated moving average (ARIMA) model fitted to the monthly data. Slightly abusing notation, let x_t be the monthly series. Using standard Box and Jenkins techniques and data from 1970 to 1984, we found the following model to be reasonable for the monthly Hawaiian tourist series: $(1-\phi B)(1-B)(1-B^{12})\log(x_t) = c + (1-\theta_1 B)(1-\theta_{12}B^{12})e_t$. We construct the prediction of the season total in a direct way. At the kth month, we use the estimated seasonal ARIMA model to obtain one-step- to $(12-k)$-step-ahead predictions for months $k+1$ to 12. The predictions are then transformed back to the original scale, with bias adjustments. See, for example, Pankratz (1983, p. 257). The predicted yearly total is then obtained by the sum of these predictions plus the observed values of months 1 to k. The prediction results are presented in Figure 5 with the symbol "o." Many linear transfer function models using the exogenous variables mentioned in Section 2.3 were investigated. None of them worked as well as the ARIMA model, partially due to the poor quality of the long-term prediction needed for the exogenous variables.

Figure 6 shows the mean squared errors and mean absolute errors of the competing procedures for the out-of-sample period 1985–1995. We can see that the SSP–Poisson AR(1)x model performs best in long-term prediction—that is, at the beginning of the year. The SSP-Gaussian AR(1) model and the ARIMA model perform equally well in short-term prediction, but the SSP-Gaussian AR(1) performs better in the intermediate term. The sum of the mean squares over all lead times indicates that the best method is SSP-Gaussian AR(1), and ARIMA is the worst. Summing the mean absolute errors over all lead times indicates that the SSP-Poisson AR(1)x is the best while the ARIMA model is the worst. Overall, in this particular study, SSP models perform better than the ARIMA model. Although not examined here, it is possible to combine the procedures, as in the work of Granger and Ramanathan (1984).

We note that the SSP model is simpler and more straightforward in providing prediction and mean squared prediction errors, but it is more difficult to do so with the ARIMA model using monthly data because it combines several multistep predictions. On the other hand, SSP has much smaller sample size in fitting the model for the season total y_t. Model selection and model checking are more difficult and nonstandard.

Finally, we remark on the differences between the discrete SSP models and the continuous SSP models. When the scale is large, count data can be treated as continuous data. In this case, both the discrete and continuous SSP models can be used. The prediction results are certainly different, as we have seen in the Hawaii series. The differences come from three sources—the estimation and prediction using the

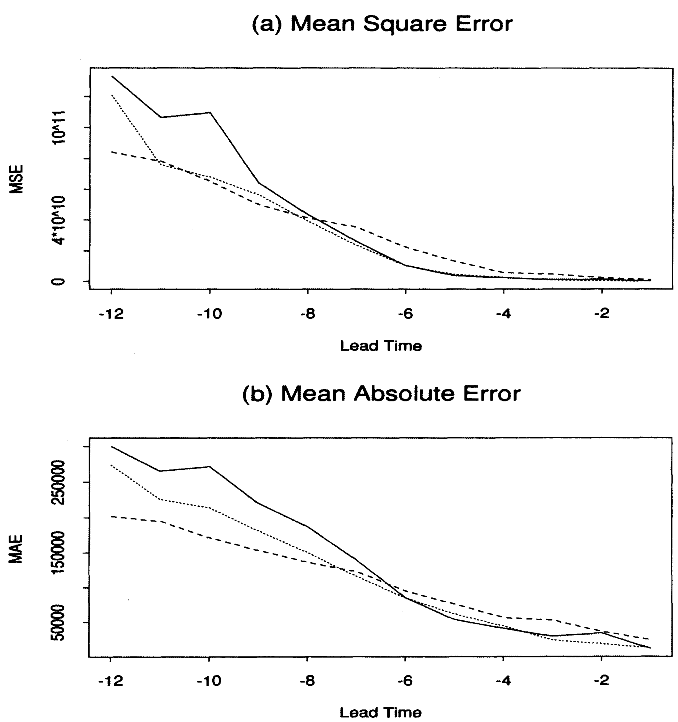

Figure 6. Mean Squared Error (a) and Mean Absolute Error (b) of Different Prediction Procedures for Out-of-Sample Period 1985–1995: ———, ARIMA; - - - -, Gaussian AR(1); ---, CPAR(1)x.

Table 3. Relative Weights of the Available Monthly Information Versus the Predicted Remaining Year Total for the G-MN and Discrete Model Over Various Horizons, year 1995

		k										
Models		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
G-MN	w_1	.56	.74	.83	.88	.91	.92	.95	.97	.98	.99	.99
c = .686	w_2	.44	.26	.17	.12	.09	.08	.05	.03	.02	.01	.01
G-MN c = 1	w_1	.08	.16	.25	.33	.40	.49	.58	.68	.76	.83	.91
Discrete	w_2	.92	.84	.75	.67	.60	.51	.42	.32	.24	.17	.09

annual series, the estimation and prediction of the monthly proportions θ_j, and the updating rule, which combines the prediction of the remaining season total ($\sum_{i=k+1}^{d} \theta_i \lambda_t$ or $\sum_{i=k+1}^{d} \theta_i \mu_t$) and the available information for the current season ($\sum_{i=1}^{k} x_{ti}$).

In the Hawaii series, the CPAR(1)x model seems to predict the unadjusted yearly total better than the Gaussian AR(1) model, using only the historical annual data (i.e., lead time -12 in Fig. 6). The monthly proportion θ's do not differ significantly between the multinomial model and the G-MN model. See Tables 1 and 2. The variances are significantly different, however. For the discrete model, the variance of monthly data, given the yearly total, is $y_t\theta_i(1 - \theta_i)$, whereas the G-MN models assumes constant variance $\sigma^2\theta_i(1 - \theta_i)$. For year 95, the ratio of the standard deviation of the discrete model to the standard deviation of the G-MN model is .04, saying that the discrete model assumes much smaller monthly variation, given the yearly total. This variance difference dictates substantially different weighting of the predicted remaining year total and the available monthly information across the two models. In the discrete SSP model, Equation (2) states that the yearly total prediction is the simple sum of the available monthly information and the predicted remaining year total. On the other hand, in the continuous SSP model, the weight depends on c, the variance ratio of the G-MN and the Gaussian yearly series, and γ_k, the amount of current monthly information available, as shown in Equation (6). Table 3 shows the weight changes for the G-MN and the discrete models over different horizons for year 95, based on all the data, 1970–1995. The $c = \sigma^2/\sigma_y^2$ value for year 95 is .0686. We can see that the weights are significantly different from the case using $c = 1$, which is used in the discrete SSP model. With $c = .0686$, the Hawaii series puts significantly larger weights on the monthly G-MN model indicating that the annual AR(1) trend model is not good and thus improvements can be made to the overall model by improving the modeling of the annual series.

ACKNOWLEDGMENTS

We thank Barbara Okamoto, Director of Market Research at the Hawaii Visitors Bureau, for generously providing the Hawaiian Tourism data to us and Leroy Laney, Chief Economist and Senior Vice President of First Hawaiian Bank, for many insightful discussions on the Hawaiian economy. We also thank the editor, an associate editor, three referees, and R. H. Shumway for their helpful comments and suggestions, which improved the manuscript dramatically. This research was supported in part by National Science Foundation grant DMS 9626113.

[Received March 1997. Revised February 1999.]

REFERENCES

Box, G. E. P., and Jenkins, G. M. (1976), *Time Series Analysis: Forecasting and Control*, San Francisco: Holden-Day.

Burman, P., and Shumway, R. H. (1998), "Semiparametric Modelling of Seasonal Time Series," *Journal of Time Series Analysis*, 19, 127–145.

Chang, S. H., and Fyffe, D. E. (1971), "Estimation of Forecast Errors for Seasonal Style-Goods Sales," *Management Science*, B, 16, 89–96.

Gilbart, J. W. (1854), "The Law of the Currency as Exemplified in the Circulation of Country Bank Notes in England Since the Passing of the Act of 1844," *Statistical Journal*, 17.

——— (1856), "The Laws of Currency in Scotland," *Statistical Journal*, 19.

——— (1865), *Logic for the Million, a Familiar Exposition of the Art of Reasoning*, London: Bell and Daldy.

Granger, C. W. J., and Ramanathan, R. (1984), "Improved Methods of Combining Forecasts," *Journal of Forecasting*, 3, 197–204.

Green, M., and Harrison, P. J. (1973), "Fashion Forecasting for a Mail-Order Company Using a Bayesian Approach," *Operations Research Quarterly*, 24, 193–205.

Grinold, R. C., and Marshall, K. T. (1977), *Manpower Planning Models*, New York: Elsevier–North-Holland.

Harvey, A. C. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge, U.K.: Cambridge University Press.

Harvey, A. C., and Fernandes, C. (1989), "Time Series Models for Count or Qualitative Observations," *Journal of Business & Economic Statistics*, 7, 407–417.

Hylleberg, S. (1992), "The Historical Perspective," in *Modelling Seasonality*, ed. S. Hylleberg, Oxford, U.K.: Oxford University Press, pp. 15–25.

Johnston, J. (1984), *Econometric Methods*, New York: McGraw-Hill.

Marshall, K. T., and Oliver, R. M. (1970), "A Constant Work Model for Student Enrollments and Attendance," *Operations Research Journal*, 18, 193–206.

——— (1979), "Estimating Errors in Student Enrollment Forecasting," *Research in Higher Education*, 11, 195–205.

McKenzie, E. (1985), "Some Simple Models for Discrete Variate Time Series," *Water Resources Bulletin*, 21, 645–650.

——— (1988), "Some ARMA Models for Dependent Sequences of Poisson Counts," *Advances in Applied Probability*, 20, 822–835.

Oliver, R. M. (1987), "Bayesian Forecasting With Stable Seasonal Patterns," *Journal of Business & Economic Statistics*, 5, 77–85.

Pankratz, A. (1983), *Forecasting With Univariate Box–Jenkins Models: Concepts and Cases*, New York: Wiley.

Tsay, R. S. (1992), "Model Checking via Parametric Bootstraps in Time Series Analysis," *Journal of Applied Statistics*, 41, 1–15.

Tweedie, R. L. (1975), "Sufficient Conditions for Ergodicity and Recurrence of Markov Chains on a General State Space," *Stochastic Processes and Their Applications*, 3, 385–403.